EG UK Computer Graphics & Visual Computing (2018)
G. Tam and F. Vidal (Editors)

Selecting Feature Representation for Online Summarisation of Egocentric Videos

Paria Yousefi[†], Ludmila I. Kuncheva and Clare E. Matthews

School of Computer Science, Bangor University, Bangor, UK

Abstract

Visualising the content of a video through a keyframe summary has been a long-standing quest in computer vision. Using real egocentric videos, this paper explores the suitability of seven feature representations of the video frames for the purpose of online summarisation. Computational speed is an essential requirement in this set-up. We found that simple feature spaces such as RGB moments and CENTRIST are a good compromise between speed and representativeness in comparison with semantically richer but computationally more cumbersome spaces obtained through convolutional neural networks.

CCS Concepts

•***Computing methodologies*** → ***Computer vision; Video summarization; Image representations;***

1 Introduction

Video summarisation is the task of representing a video by a small and informative set of frames (keyframes) from the video [TV07, MA08,MTLT17]. The lack of clear structure and visual consistency of first-person videos (FPV), called also egocentric videos, make their summarisation substantially more difficult [Bam15,MTLT17, BDR17,BMT*15].

Adding to the challenge, here we are interested in online summarisation of egocentric videos. In online summarisation, the keyframe summary is built on-the-fly so that at any moment during the video capture, there is a valid summary of the video up to that moment. While studies on online video summarisation exist [AA08, OLS*15, RS03, MGW*15, AMT16, EK17], none is specifically dedicated to egocentric videos. One important aspect of the video summarisation pipeline is the extraction of features from the video frames. Ideally, the feature representation will capture both the semantic content and the visual appearance of the frame. Many such representations have been proposed in the literature, ranging from low-level features (e.g., colour spaces) to high-level features (semantic-level description of the image content). In this paper, we examine the suitability of seven feature spaces for online summarisation of egocentric video. Our experimental analysis is based on egocentric videos from an activity recognition database [PR].

The rest of the paper is organised as follows. Section 2 describes the feature spaces. Section 3 contains our experiments, and Section 4, our conclusions and future work.

2 Feature representations

For an online application, two factors must be considered when choosing a descriptor: (1) the ability of the chosen feature space to identify the meaningful attributes of the scene; (2) the computational cost of processing (the extraction process, and algorithm running time associated with the feature dimensionality).

In order to select an appropriate feature space for an online algorithm, we analyse a number of different features (detailed in Table 1):

1. *RGB moments.* The RGB colour moments are obtained by dividing an image uniformly into 3×3 blocks. The mean and the standard deviation for each block and colour channel are computed.
2. *Colour Layout (MPEG7).* [KY01] An input RGB image is uniformly divided into 8×8 blocks. The average value of the pixel colours for each block is calculated. The average RGB colours is converted into YCbCr colour space and then quantized into three sets of 64 DCT coefficients (total of 192 features).
3. *CENTRIST descriptor.* CENsus TRansform hISTogram (CENTRIST) [WR11]. Census Transform compares the intensity value of a pixel with its eight neighboring pixels. The binary results from the 8 comparisons are transformed in a decimal number between 0 and 255. A histogram of these numbers is then generated with 256 bins, one for each Census intensity. The two end bins (corresponding to 0 and 255) are removed, leaving a 254-dimensional feature space. We used a MATLAB implementation to extract the descriptor [Boe].
4. *HSV histograms.* The feature space is extracted by a quantisation of the HSV color space into a 256-dimensional histogram vector of 32 bins for Hue, 4 bins for Saturation and 2 bins for Value.

[†] Corresponding author. E-mail: paria.yousefi@bangor.ac.uk

To increase speed the original image is resized to 1/64th of its original size.

5. *GIST.* [OT01] The Gist descriptor is computed by convolving an image with 32 Gabor filter (4 scales and 8 orientations), producing 32 feature maps. Each feature map is divided into 4 × 4 regions and the average feature values calculated for each region. The 16 average values of 32 feature maps are concatenated resulting 512-dimensional descriptor.
6. and 7. *Places205-AlexNet and VGGNet.* We included two high level feature descriptors extracted through deep learning neural networks. The 4096 deep features are extracted right before the classification (soft-max) layer of two pre-trained Convolutional Neural Networks (CNNs), known as: VGGNet architecture [SZ14] available through the MatConvNet toolbox [VL15]; and Places205-AlexNet model [ZLX*14] using Caffe deep learning toolbox [JSD*14].

3 Experiment

The purpose of the experiment is to evaluate the feature spaces in regard to their suitability for online keyframe summarisation from egocentric video. Thus, we consider two aspects: ease of calculation of the feature space and the quality of the produced summary.

We chose the Activity of Daily Living (ADL) dataset [PR12]. The ADL dataset was recorded using a chest-mounted GoPro camera which consists of 20 videos of subjects performing their daily activities in the house.

Table 1: *Comparison of the average time of feature extraction for the toy video and the average MCC-value for all 20 videos.*

		Image Size		Visual Info.					
	Used in	Resized	Original	Colour	Scene	Deep learning	Dimensions	Time(sec)	MCC-value
RGB moments	[MYK18]		✓	✓			**54**	50	**0.68**
Color Layout	[OLS*15]		✓	✓			192	519	0.52
CENTRIST	[MGW*15]		✓		✓		254	160	0.63
HSV histogram	[ALT12]	✓		✓			256	**30**	0.45
Gist	—	✓			✓		512	232	0.45
Places205-AlexNet	—		✓		✓	✓	4096	494	0.46
VGGNet	[AMT16]		✓			✓	4096	2377	0.43

3.1 Extraction time

All experiments were carried out on a laptop, 2.20 GHz Intel Core *i*5 CPU, with 8GB RAM. The first part of our analyses compares the processing time to extract the different features for the toy video. The 'toy video' is a selection of the initial 495 frames from video #8 of the same dataset. For each descriptor, we calculated the average time of extraction by repeating the process 20 times. The results are shown in Table 1. The extraction time for the simple colour spaces (RGB moments and HSV histograms) is shorter than the time for the other descriptors, whereas the popular VGGNet has the longest extraction time.

3.2 Quality of the keyframe summary

The second part of our analyses compares the qualities of the summaries based on the different feature spaces.

3.2.1 The online summarisation algorithm. We used an online summarisation algorithm based on control-charts [MYK18]. The algorithm monitors the distance between consecutive frames (points in the chosen feature space), and detects a transition between events in the video when this distance exceeds a threshold. The algorithm starts with an initial buffer of frames, B. The mean, μ and standard deviation, σ of distances between consecutive frames in B is calculated. While the distance between subsequent consecutive frames is less than the threshold of $\mu + 3\sigma$, frames are added to B and μ and σ updated. A distance greater than the threshold identifies the end of an event. If B is larger than a specified minimum, the frame closest to the centroid of the event is selected (otherwise the event is ignored). This frame is compared to the previously selected keyframe (if it exists), and it is only added to the summary S if it is sufficiently different. The similarity of the selected frames is measured by the Pearson correlation coefficient. If the two frames are too similar, their respective events are merged, and a new, single representative keyframe selected. For each video and feature space, the parameter values producing the best result are used.

3.2.2 Performance measure. We chose the Matthews correlation coefficient (MCC) [Mat75] between the selected summary S and a given ground truth as a performance indicator. The ground truth for the dataset was created as follows: Each event in the video is distinguished by a number of terms. The frames in an event are labelled as informative/not informative based on whether they contain semantic information that is included in the relevant terms for this event. Consequently, any informative frame from the event can be considered ground truth for that event.

3.2.3 Results. The average MCC-values using the chosen feature space for 20 videos are shown in Table 1. The higher the value, the better the quality of the summary. The RGB moments has the highest MCC-value, and the VGGNet descriptor, the lowest value. CENTRIST feature space gave better performance than CNN, and was also faster to extract. The difference between MCC-values for the HSV histogram, Gist and the CNN descriptors are not large. However, the HSV histogram has fewer dimensions and substantially faster processing time.

4 Conclusions

The experiments show that for our online summarisation of egocentric videos, simple, colour-based descriptors offer a substantially more efficient and higher quality summary than the complex CNN features tested. For the colour-based descriptors, the use of resized images does not appear to adversely affect the summary quality. Image compression is therefore an interesting area to explore for online video summarisation, with a potential for further gains in efficiency.

Acknowledgement

This work was done under project RPG-2015-188 funded by The Leverhulme Trust, UK.

References

[AA08] Abd-Almageed W.: Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing. In *IEEE 15th International Conference on Image Processing (ICIP 2008)* (Oct. 2008), pp. 3200–3203. 1

[ALT12] Almeida J., Leite N. J., Torres R. d. S.: Vison: Video summarization for online applications. *Pattern Recognition Letters 33*, 4 (Mar. 2012), 397–409. 2

[AMT16] Anirudh R., Masroor A., Turaga P.: Diversity promoting online sampling for streaming video summarization. In *IEEE International Conference on Image Processing (ICIP2016)* (Sept. 2016), pp. 3329–3333. 1, 2

[Bam15] Bambach S.: A survey on recent advances of comp. vision algorithms for egocentric video. *arXiv:1501.02825* (2015). 1

[BDR17] Bolaños M., Dimiccoli M., Radeva P.: Towards storytelling from visual lifelogging: An overview. *Journal of Transactions on Human-Machine Systems 47* (2017), 77–90. 1

[BMT*15] Bolaños M., Mestre R., Talavera E., Giró i Nieto X., Radeva P.: Visual summary of egocentric photostreams by representative keyframes. In *Proc. IEEE Int. Multimedia and Expo Workshops (ICME)* (2015), pp. 1–6. 1

[Boe] Boehm S.: Matlab centrist. https://github.com/sometimesfood/spact-matlab. Accessed: 2018-08-01. 1

[EK17] Elhamifar E., Kaluza M. C. D. P.: Online summarization via submodular and convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)* (Jul. 2017), pp. 1818–1826. 1

[JSD*14] Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell T.: Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (2014), MM '14, pp. 675–678. 2

[KY01] Kasutani E., Yamada A.: The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Proceedings 2001 International Conference on Image Processing (ICIP01)* (October 07-10 2001), vol. 1, IEEE, pp. 674–677. 1

[MA08] Money A. G., Agius H. W.: Video summarization: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation 19*, 2 (2008), 121–143. 1

[Mat75] Matthews B. W.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure 405*, 2 (1975), 442–451. 2

[MGW*15] Mei S., Guan G., Wang Z., Wan S., He M., Feng D. D.: Video summarization via minimum sparse reconstruction. *Pattern Recognition 48*, 2 (Feb. 2015), 522–533. 1, 2

[MTLT17] Molino A. G. D., Tan C., Lim J. H., Tan A. H.: Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems 47*, 1 (2017), 65–76. 1

[MYK18] Matthews C. E., Yousefi P., Kuncheva L. I.: Using control charts for on-line video summarisation. In *14th Asian Conference on Computer Vision (ACCV 2018)* (July 2018). Submitted. 2

[OLS*15] Ou S.-H., Lee C.-H., Somayazulu V. S., Chen Y.-K., Chien S.-Y.: On-line multi-view video summarization for wireless video sensor network. *IEEE Journal of Selected Topics in Signal Processing 9*, 1 (Feb. 2015), 165–179. 1, 2

[OT01] Oliva. A., Torralba A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. Journal of Computer Vision 42*, 3 (2001), 145–175. 2

[PR] Pirsiavash H., Ramanan D.: ADL dataset. https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/. Accessed: 2018-08-01. 1

[PR12] Pirsiavash H., Ramanan D.: Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR12)* (June 16-21 2012), IEEE, pp. 2847–2854. 2

[RS03] Rasheed Z., Shah M.: Scene detection in hollywood movies and tv shows. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Jun. 2003), vol. 2, pp. 343–343. 1

[SZ14] Simonyan K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 2

[TV07] Truong B. T., Venkatesh S.: Video abstraction. *ACM Transactions on Multimedia Computing, Communications, and Applications 3*, 1 (2007), 3–es. 1

[VL15] Vedaldi A., Lenc K.: Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia* (2015). 2

[WR11] Wu J., Rehg J. M.: Centrist: A visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 33*, 8 (2011), 1489–1501. 1

[ZLX*14] Zhou B., Lapedriza A., Xiao J., Torralba A., Oliva A.: Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems 27* (2014), Ghahramani Z., Welling M., Cortes C., Lawrence N. D., Weinberger K. Q., (Eds.), Curran Associates, Inc., pp. 487–495. 2